Neurocomputing 613 (2025) 128691

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

FedATA: Adaptive attention aggregation for federated self-supervised medical image segmentation

Jian Dai [a], Hao Wu [c], Huan Liu [b], Liheng Yu [a], Xing Hu [d], Xiao Liu [b], Daoying Geng [a,b,*]

[a] Academy for Engineering and Technology, Fudan University, Shanghai, 200433, China
[b] Department of Radiology, Huashan Hospital, Fudan University, 200040, China
[c] Department of Dermatology, Huashan Hospital, Fudan University, 200040, China
[d] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, 20093, China

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Self-supervised learning
Masked image modeling
Federated learning

ABSTRACT

Pre-trained on large-scale datasets has profoundly promoted the development of deep learning models in medical image analysis. For medical image segmentation, collecting a large number of labeled volumetric medical images from multiple institutions is an enormous challenge due to privacy concerns. Self-supervised learning with mask image modeling (MIM) can learn general representation without annotations. Integrating MIM into FL enables collaborative learning of an efficient pre-trained model from unlabeled data, followed by fine-tuning with limited annotations. However, setting pixels as reconstruction targets in traditional MIM fails to facilitate robust representation learning due to the medical image's complexity and distinct characteristics. On the other hand, the generalization of the aggregated model in FL is also impaired under the heterogeneous data distributions among institutions. To address these issues, we proposed a novel self-supervised federated learning, which combines masked self-distillation with adaptive attention federated learning. Such incorporation enjoys two vital benefits. First, masked self-distillation sets high-quality latent representations of masked tokens as the target, improving the descriptive capability of the learned presentation rather than reconstructing low-level pixels. Second, adaptive attention aggregation with Personalized federate learning effectively captures specific-related representation from the aggregated model, thus facilitating local fine-tuning performance for target tasks. We conducted comprehensive experiments on two medical segmentation tasks using a large-scale dataset consisting of volumetric medical images from multiple institutions, demonstrating superior performance compared to existing federated self-supervised learning approaches.

1. Introduction

The accurate and robust medical image segmentation [1] play a crucial role in facilitating accurate diagnosis, early detection, and treatment plan optimization. For instance, automatically and rapidly identifying and measuring lesions from MRI/CT scans is a vital step in preventing diseases and optimizing treatment plans. However, the model experiences a noticeable decrease in performance due to a lack of large-scale datasets for training. These sensitive and private medical datasets are usually stored in isolated medical institutions, making it impractical and illegal to construct a centrally large-scale medical datasets. One efficient solution to overcome this obstacle is to utilize federated learning, which trains the model in a distributed manner across multiple institutions by exchanging model data rather than raw medical image data. Recent works [2–5] have demonstrated that federated learning provides a promising solution to training shared models on distributed datasets without considering privacy concerns.

Conventional federated learning methods collaboratively learn the global model among institutions through supervised learning, which requires accurate annotation of medical volumetric images in each institution. To the best of our knowledge, the expertise requirements and time-consuming are considerable barriers to annotating all the medical volumetric images, making it challenging to train an accurate segmentation model relying only on limited labeled data. Self-supervised learning [6–9] has recently gained attention that pre-train in proxy tasks and fine-tuning for downstream models with limited labeled data. Therefore, self-supervised federated learning (FSL) is proposed to combine self-supervised learning with federated learning for medical

* Corresponding author at: Academy for Engineering and Technology, Fudan University, Shanghai, 200433, China.
E-mail address: daoyinggeng@fudan.edu.cn (D. Geng).

https://doi.org/10.1016/j.neucom.2024.128691
Received 20 May 2024; Received in revised form 9 September 2024; Accepted 29 September 2024
Available online 22 October 2024

image segmentation. In FSL, each institution first utilizes its unlabeled data to learn its own feature space and then collaboratively aggregates these feature representations among institutions.

Existing FSL methods [10–12] have completed target segmentation tasks in medical image analysis. Yan et al. [11] propose Fed-MAE to learn the intrinsic features with unlabeled data and transfer representation to the downstream tasks by fine-tuning with limited labeled data. Wu et al. [12] propose a novel self-supervised method to integrate contrastive learning with federated learning for generic representation learning through volumetric image argument operation. Although these methods improve the performance for further fine-tuning, the descriptive capability of the pre-trained model was poor due to the existing representation gap, such as specific morphological features of lesions between pre-train and downstream tasks. We argued that (1) masked image modeling directly restores raw pixels as low-level reconstruction targets, failing to preserve contextual information specifically for high semantic-level tasks. (2) Contrastive learning only generates representation for local salient regions through central crop augmentation. Considering these issues, we introduced mask self-distillation to produce the latent representation as reconstruction targets, aiming to improve the high descriptive capability of the model and thus further eliminate the representation gap. The resulting objective is a self-distillation paradigm where the student model, similar to the MAE structure, consists of the asymmetric encoder-decoder. In contrast, the teacher model contains an encoder to produce latent representation and updates weights from the student branch using Exponential Moving Average (EMA). The knowledge is distilled from the whole image (fed to the teacher model) to the masked image (provided to the student model).

Another challenge in self-supervised federated learning is heterogeneity data distribution among multiple institutions. Regarding data heterogeneity, for instance, larger institutions have more detailed patient data than small institutions due to diverse patient populations and disease manifestation (i.e., statistical heterogeneity); some institutions collect and annotate interest organs due to the various interests in clinical studies (i.e., label heterogeneity). Each institution only applies mask self-distillation on local data to produce its feature representation while not considering heterogeneity data distribution among institutions. This result in inconsistent feature representation across institutions. When using the aggregated model by FedAvg [13] directly, it generally suffers from generalization deterioration due to the global model including desired and undesired representation information derived from inconsistent space. However, only the desired task-related representation information is beneficial for fine-tuning locally. If using all the knowledge of the global model, the global pre-trained model can not provide accurate and related knowledge for individual clients. Thus, we intend to adaptively aggregate the global and local models toward refocusing specific-related representations to local fine-tuning for higher task performance. Since the self-attention layer is rooted in the self-attention mechanism and effectively learns feature representation ability more than other layers, we preform aggregation on the self-attention layer in an adaptive way and further reduce the computation overhead. Therefore, we proposed Adaptive Attention Aggregation with federated learning (FedATA) as partial personalized FL, allowing the global model to refocus its attention on task-related representation through personalized attention layers.

In this work, we propose a novel self-supervised federated learning method that combines masked self-distillation with personalized federated learning to facilitate segmentation performance under data heterogeneity and label deficiency. In particular, we introduce masked self-distillation which utilizes high-quality latent representations as targets, effectively bridging the representation gap in downstream segmentation tasks. Furthermore, we propose adaptive attention aggregation with personalized federated learning to precisely capture the desired representation for fine-tuning locally. We comprehensively validate our proposed framework in CT/MRI scenarios using five public datasets, and the results show the remarkable potential of FedATA for medical image segmentation tasks.

Our contributions can be summarized as follows:

1. We propose a novel distributed self-supervised learning framework for medical image segmentation, which learns visual representation from decentralized institutions, followed by transferring specific-related knowledge for personalized fine-tuning.
2. We introduce a mask self-distillation method to eliminate the representation gap by constructing a high-level reconstruction target, rather than the low-level pixel of the volumetric image.
3. we propose adaptive attention aggregating with federated learning that adaptively aggregates the global and local models to capture specific-related representation and further boost local fine-tuning performance.
4. we empirically demonstrate our FedATA method on five public volumetric medical image segmentation datasets. Our experimental result show consistent performance improvement across different scenarios.

2. Related work

2.1. Self-supervised learning

The attention on self-supervised visual learning has risen over the past few years. The core of self-supervised learning is adopting generic representation learning in pretext tasks and then fine-tuning for target downstream models. The self-supervised learning approach is partitioned into contrastive learning [14–19] and masked image modeling [20–23]. Contrastive learning applies contrastive loss to compute the similarity and dissimilarity with the target sample between two or more views. For example, BYOL [15] trains the network to predict the representation of the same image under different augmented views obtained from the other network. Therefore, contrastive learning depends on enough data argument operation and takes up much computation space in medical image analysis. Zhang et al. [19] demonstrate that contrastive learning can improve the precision of medical image classification in mix-up domains. However, contrastive learning focuses on the central region, dismissing global representation for fine-tuning downstream tasks. Masked image modeling methods show more promising fine-tuning performance by reconstructing the remaining masked portion of its original patch from the visible portion. Zhou et al. [22] demonstrate that the masked image modeling approach can significantly promote volumetric medical image analysis and accelerate the convergence of supervised learning compared to contrastive learning. Yan et al. [11] apply masked image modeling in diverse medical imaging tasks and facilitate effective representation learning with unlabeled data.

2.2. Knowledge distillation

Self-knowledge distillation [24–26] is a simple yet effective approach that aims to distill the knowledge from itself and make full use of itself for training the robust model. An effective way is transferring the valuable knowledge of the teacher network to the student network by using the output of the teacher network as the objective for training the student. Different from distilling knowledge from the pre-trained teacher model, the student structures are equivalent to the teacher, and the student model's temporal ensemble gradually becomes the teacher model in the self-knowledge distillation. The benefit of self-knowledge distillation is gradually utilizing the teacher model's knowledge to soften the challenging target and capture more useful information while training the whole network. Recently, the self-knowledge distillation paradigm has also been investigated in semi-supervised learning [27], contrastive learning [28], and mask image modeling [29]. This paper uses masked self-distillation to produce a latent prediction target for guidance student model learning, which naturally fits masked image modeling to capture more

transferable knowledge.

2.3. Federated learning

Federated learning [30,31] is a distributed machine learning approach for training privacy-preservation models through aggregating local models instead of local raw data, widely applied in medical image analysis. The known traditional federated learning method is FedAvg [13], which optimizes the objective by performing stochastic gradient descent updates at each client and average aggregation on the server side. However, FedAvg generally fails to achieve better performance for all clients due to generating inconsistent feature space by the data heterogeneity. Many studies have attempted to address these issues through personalized federated learning, which focuses on training a client-specific model instead of a unified global model. Sun et al. [30] propose a personalized federated learning framework with partial model to account for the diversity of data distributions across different clients. Feng et al. [32,33] propose an encoder-decoder structure within a federated learning framework for magnetic resonance image reconstruction. A globally shared encoder is maintained on the server and then learns domain invariant representation. At the same time, a client-specific decoder is trained with local data to take advantage of the domain-specific properties of each client.

3. Method

3.1. Overview of framework

Our work aims to train a robust segmentation model that learns from distributed clients with limited annotation and heterogeneity data. Specifically, our goal is to enhance the quality of the pre-trained model and facilitate local fine-tuning for volumetric medical image segmentation. The design of our proposed method is illustrate in Fig. 1. The first phase of our method involves federated self-supervised pre-training, followed by the supervised federated fine-tuning phase. Since the supervised fine-tuning can be completed with little effort using available annotations, our efforts focus on learning the pre-trained model in the self-supervised stage.

During the federated self-supervised stage, an unlabeled volumetric image in the client is partitioned into a set of non-overlapping image patches, and the image patches are simultaneously grouped into the visible patch and the masked patch set. The teacher branch contains the encoder that feeds into all image patches to produce the latent representation that acts as a learning target. In contrast, the student branch incorporates the asymmetric encoder-decoder architecture, and the encoder in the student branch only takes the visible patches as input, which learns the contextualized representation ability. Since directly sharing raw images is prohibitive, image representations are converted to feature vectors through student encoder. As shown in Fig. 1, each client uploads its local student encoder to a server, generating a global model by aggregating the uploaded models. Client i then downloads the global model from the server and adaptively aggregates it with its existing old local model as partial model personalization to capture specific-related representations and facilitate local fine-tuning for downstream segmentation tasks. During the supervised federated fine-tuning stage, the encoder of segmentation model is initialized by the final pre-trained model, and all attention layers of the downstream model perform adaptive aggregation in FL, similar to the federated self-supervised.

3.2. Mask image modeling with self-distillation

The mask image modeling (MIM) integrates the mask self-distillation paradigm to perform representation learning in each local client, where the student network $M_s$follows the decoupled encoder-decoder transformer architecture similar to MAE [8] and the teacher network M_t only retains encoder M_t^e updated by exponential moving average (EMA) of the student network parametersM_s. The input volumetric $X_{img} \in \mathbb{R}^{T\times H\times W\times D}$ is first divided into several non-overlapping 3D patches, followed by a linear projection layer map to visual tokens.

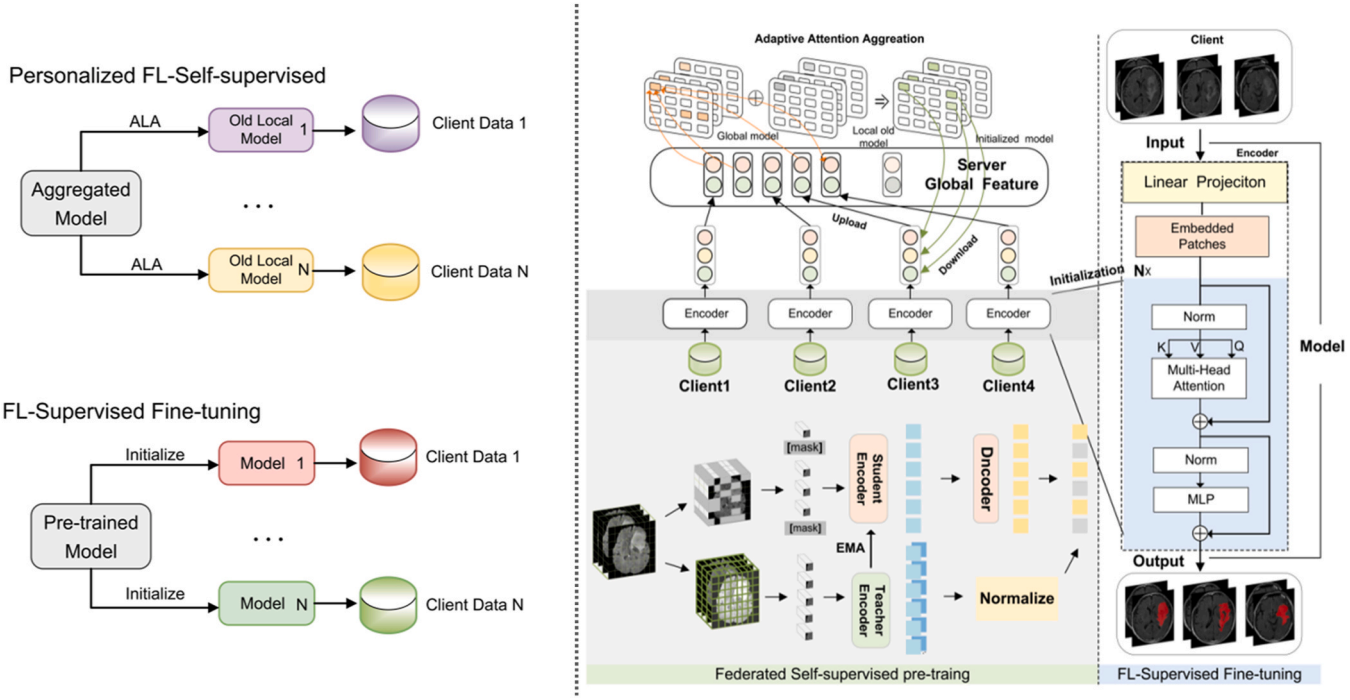

Fig. 1. The overview of the proposed Fed-ATA method. In federated pre-training, mask self-distillation is used as the self-supervised task to learn representation from unlabeled images on each client. Specifically, client i uploads its student encoder to the server. Once the server aggregates the uploaded model, the client i downloads the global model and performs adaptive aggregation with the old local model for initialization. In the FL fine-tuning, the final pre-trained model from the first stage is used to initialize the model encoder. End-to-end federated fine-tuning is then performed on labeled images for image segmentation in each local client.

Specifically, input visual tokens $\{x_i^p\}_{i=1}^N$ are fed to the teacher encoder M_t^e to obtain high-level latent representation space $\{e_i\}_{i\in N}$, where $e_i = M_t(x_i^p \text{for} i \in N)$.The student network M_s samples a random binary mask, where mask tokens $\{x_i^p\}_{i\in M}$ need to be dropped from the token sequence and only visible tokens $\{x_i^p\}_{i\in V}$ fed into encoder M_s^e. Furthermore, the student decoder M_s^d is composed of multiple Transformer blocks with multi-head self-attention layers. To reconstruct the representation $\{h_i\}_{i\in M}$ of the masked patches from visible patches in the encoded representation space and then map the predicted representation to the target high-level latent representation space $\{e_i\}_{i\in N}$, we first concatenate the visible and learned mask tokens, followed by sending the combined token sequence into the student decoder $M_s^d$to produces the output features. Subsequently, a liner layer W is applied on output features to align with the channel dimension of $\hat{Y}_i^t$ and generate prediction Y_i^s, i.e.,$Y_i^s = WM_s^d(h_i^s)$, and then training the student and teacher model to minimize distillation loss$\mathscr{L}$D that calculates the distance between the target features of masked patches and predicted features, which is shown as:

$$\mathscr{L}\mathrm{D} = \sum_{i\in\{V\cup M\}} \mathrm{D}(\hat{Y}_i^t, \quad WM_s^d(h_i^s)) \tag{1}$$

Masked Self-Distillation: The mask self-distillation paradigm makes full use of the teacher modelM_t to generate latent contextualized representations instead of low-level pixels. These high-level representations, serving as targets of student networkM_s prediction tasks, are generated by teacher encoder incorporated information from the entire volumetric sample. Therefore, the training task is for the student network based on the masked version of the tokens to regress representation targets. In general, the latent representation is generated by calculating the average of the top K Transformer blocks in the teacher model. The weights Δ of teacher encoder M_t^e are an exponentially moving average of the student encoder weights θ, i.e. $\Delta \leftarrow \tau\Delta + (1-\tau)$ where τ follows a linearly increasing schedule from a staring value τ_0 to a final value τ_e over τ_n updates, after which the value is kept constant.

Distillation Loss function: when the student model captures the knowledge from the teacher model, the alignment constraint is imposed on the latent representation$\hat{Y}_i^t$from the teacher encoder and the representation Y_i^s from the student decoder. We use the distillation loss to align the two latent representations $\hat{Y}_i^t$ and predicted representations Y_i^s. The distillation loss is defined as:

$$\mathrm{D} = \mathscr{L}_1(\mathrm{Norm}(\hat{Y}_i^t), \quad Y_i^s) \tag{2}$$

where Norm(.)is the feature scaling operation performed by calculating the mean and standard deviation of the target representation from teacher encoder. $\mathscr{L}_1$ is the smooth L1 loss between L1 and L2 loss:

$$\mathscr{L}_1(y, \quad \hat{y}) = \begin{cases} \frac{1}{2}(\hat{y}-y)^2 \Big/ \beta, & |\hat{y}-y| \le \beta \\ |\hat{y}-y| - \frac{1}{2}\beta, & \text{otherwise} \end{cases} \tag{3}$$

where β is set to 1.5 in our experiment.

Algorithm 1. Mask self-distillation learning in PyTorch-like style.

```
# f: student encoder
# t: teacher encoder
# g: decoder for target features
# t_m: learnable mask tokens
# Fd_loss: distillation loss
# EMA: exponential moving average
for epoch in epochs:
    # x: volumetric data, m: mask
    for x, m in loader:
        # patch-based input embedding
        x_pre = patch_emb(x)
        # masking tokens
        x_vis = select_mask(x_pre, 1 - m)
        # visible local patch features
        q_vis = f(x_vis)
        # reconstruction target features
        p_img = g(concat(q_vis, t_m))
        # compute target feature
        t_img = t(x_pre)
        # compute reconstruction loss
        loss = Fd_loss(p_img, k_img)
        # loss backward
        loss.backward()
        # student and teacher parameter update
        t_parameter = EMA(f_parameter)
        # optimizer update
        optimizer.step()
```

3.3. Attention adaptive aggregation with FL

In the previous section, we introduced mask self-distillation to improve the performance of the pre-trained model in each client. Next, preserving the generalization of the pre-trained model across all clients is crucial, especially when dealing with heterogeneous data distributions. Traditional FedAvg [13] for model aggregation is often suboptimal as it assigns the same parameters to all clients regardless of the unique data distributions. In contrast, personalized federated learning (PFL) approaches, such as FedPerfix [30], are particularly effective in environments with heterogeneous data. As shown in Fig. 2, we compare the attention maps of three federated learning methods to understand how each method focuses attention during self-supervised learning. Our

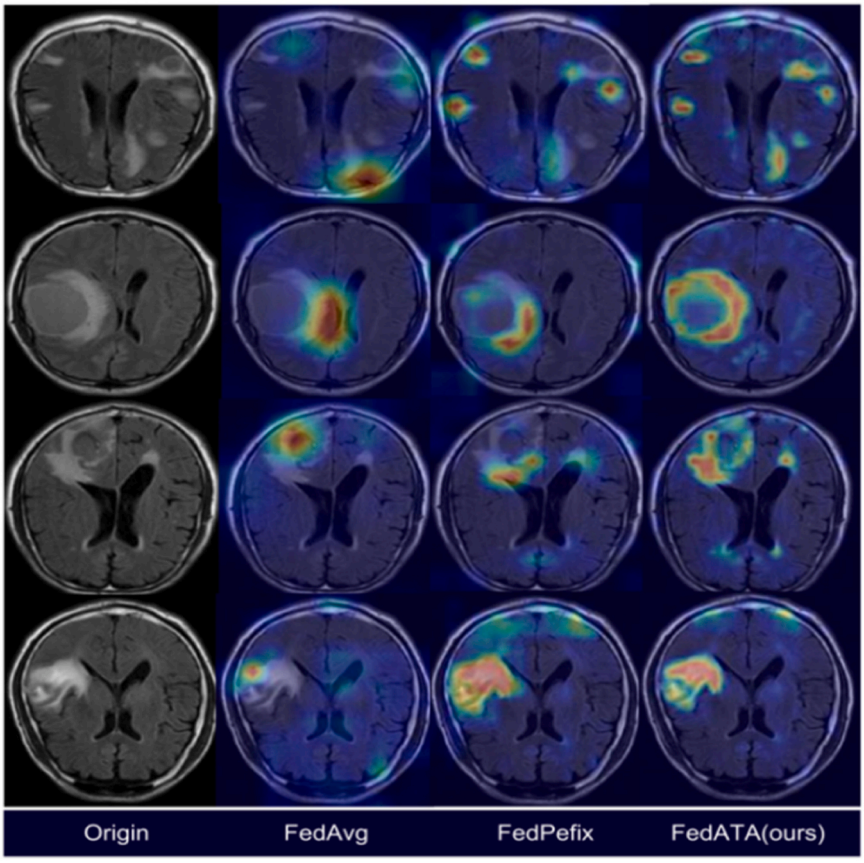

Fig. 2. Attention maps of FedATA and other federated learning methods. We use the output of the second self-attention layer of encoder to produce attention map.

method FedATA as a form of PFL differs from FedPerfix in that it does not require additional plugin parameters. It comprises of three key operations: 1) Local Training in clients; 2) Model Communication; 3) Model adaptive attention aggregation. We assume that there are N clients, each client i with a local dataset D_i consisting of $m_i$samples drawn from a distinct data distribution. Let $D=\bigcup_{i\in N} D_i$ denote the total datasets with the size of $M=\sum_{i=1}^{N} m_i$. Let $f(\phi_i)$ denote local model for client i, parameterized by ϕ_i.The optimization objective is:

$$\operatorname{argmin}\sum_{i=1}^{N}\frac{m_i}{M}\mathscr{L}_1 f(\phi_i) \tag{4}$$

Local Training First, each client i conducts self-supervised training on unlabeled data D_i. Regardless of our FedATA method, self-attention layers are rooted in multi-head attention mechanisms and exhibit stronger representation learning capabilities compared to other layers. Inspired by this, we train local model by learning personalized self-attention layers, which effectively reduces computation overhead and mitigates privacy concerns. The queries, keys, and values in the self-attention layer are denoted as follows: $\mathrm{Q}=\mathrm{HW}^{\mathrm{Q}},\mathrm{K}=\mathrm{HW}^{\mathrm{K}}$ and $\mathrm{V}=\mathrm{HW}^{\mathrm{V}}$.The three projection parameters concatenated as$\mathrm{w}=[\mathrm{W}^{\mathrm{Q}},\mathrm{W}^{\mathrm{K}},\mathrm{W}^{\mathrm{V}}]$. Self attention layer is applied through $\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V})=\mathrm{softmax}(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d}})\mathrm{V}$, where d is the number of columns for queries, keys, and values.

Model Communication After local training, client i uploads the trained student network with self-attention layer to the server and updates it with the aggregated model. In traditional FedAvg [13], the client downloads the global model from the server and then directly replaces the local model for the next iteration, where all the parameters in one client model are assigned the same weight without considering the local objective. Considering the difference of FedAvg, we first computes the average parameter in server and sends the client i to the global parameters. The model parameter ϕ_i is be split into two parts: $\phi_i=\{\theta_i,\xi_i\}$. Let θ_i denote the projection parameters of the self-attention layer and the parameters of other layers as ξ_i. At the beginning of each round on client i, θ_i^t as personalized layer parameter need to retain in local datasets and ξ_i^t is initialized by global parameter ξ^{t-1}.

Adaptive Attention Aggregation when the client download global model from serve, it implements on the old parameter $\theta_i^{t-1}$and global parameter θ^{t-1} to obtain the optimized parameters $\widehat{\theta}_i^t$ for initialization. Instead of straightly initializing the average parameter to the local $\widehat{\theta}_i^t$ and finding the optimal $\widehat{\theta}_i^t:=\ \theta^{t-1}$ for the next iteration, we use an adaptive learning method to learn personalized parameter. Formally,

$$\widehat{\theta}_i^t:=\theta_i^{t-1}\odot w_{i,1}+\ \ \theta^{t-1}\odot w_{i,2} \tag{5}$$

$w_{i,1}$ and $w_{i,2}$ are global and local parameter's aggregating weights respectively. However, it is hard to learn $w_{i,1}$ and $w_{i,2}$ with the constraint through the gradient-based learning method. Thus, we combined $w_{i,1}$ and $w_{i,2}$ by viewing Eq. (6) as:

$$\widehat{\theta}_i^t:=\theta_i^{t-1}\ \ +(\theta^{t-1}-\theta_i^{t-1})\ \ \odot w_{i,p} \tag{6}$$

where we call them $(\theta^{t-1}-\theta_i^{t-1})$ as the "update". The weight $w_{i,p}$ has the same shape as the remaining θ^{t-1}. We initialize the value of each element in $w_{i,p}$ to one in the begging and learn $w_{i,p}$ based on old $w_{i,p}$ in each iteration. Client i trains through the gradient-based learning method until converge in local datasets:

$$w_{i,p}\leftarrow w_{i,p}-\eta\nabla_{w_{i,p}}\ell(\widehat{\theta}_i^t;\theta^{t-1}) \tag{7}$$

where η is the learning rate for weight learning and set 0.001 in our

experiment.

Algorithm 2. Adaptive attention aggregation N - clients, ϕ^0 - initial global model, T- number of communication rounds, η - learning rate of local update, E - number of local epochs.

Server :
sends ϕ^0 to all clients $C^n \subset \{1,\ldots,N\}$ to initialize local models
for each communication rounds $t \in \{1,\ldots,T\}$ **do**
for each clients $C^n \subset \{1,\ldots,N\}$ **do**
$\phi_{t+1}^n \leftarrow \text{ClientUpate}(n, \phi_t)$
$\phi_{t+1} \leftarrow \sum_{n=1}^{N} \frac{m_i}{M} \phi_{t+1}^n$

ClientUpdate(n, ϕ)**:**
for each client $i \in C^n$ **do**
if $t = 1$:
clients are initialized by ϕ^0
for each local epoch $e \subset \{1,\ldots,E\}$ **do**
for batch $b \in \beta$ **do**
$\phi_i^t \leftarrow \phi_i^0 - \eta \nabla l(\phi; b)$
return ϕ_i^t to server
else $t > 1$:
client C^t locally train $w_{i,p}$ in θ_i^t by Equation (6) and fix other parameter ξ^t
while $w_{i,p}$ does not converge **do**
$w_{i,p} \leftarrow w_{i,p} - \eta \nabla_{w_{i,p}} \ell(\hat{\theta}_i^t; \theta^{t-1})$
$\hat{\theta}_i^t := \theta_i^{t-1} + (\theta^{t-1} - \theta_i^{t-1}) \odot w_{i,p}$
$\phi_i^t = \{\hat{\theta}_i^t, \xi^t\}$ as local initialize parameters
for each local epoch $e \subset \{1,\ldots,E\}$ **do**
for batch $b \in \beta$ **do**
$\phi_i^t \leftarrow \phi_i^t - \eta \nabla l(\phi; b)$
return ϕ_i^t to server

4. Experiment

In this section, we present experiments on volumetric medical image segmentation to assess the applicability of our method. We first provide detailed information regarding the datasets and experimental setup, followed by evaluating the robustness of our method across two scenarios utilizing five public datasets. Furthermore, we analyze the generalization ability of the proposed method on an out-of-distribution test dataset and evaluate its label efficiency through fine-tuning with different fractions of the annotation. Finally, we compare the performance of our method to state-of-the-art (SOTA) self-supervised federated learning methods.

4.1. Datasets

Multi-organ CT datasets. The CT dataset includes four public abdominal CT image datasets, which were created to evaluate the performance of our method. It comprises medical segmentation decathlon datasets1 (Task #9), liver tumor segmentation challenge datasets2 (LiTS), medical segmentation decathlon datasets1 (Task #7), and Beyond the Cranial Vault datasets3 (BTCV). We refer to each dataset as the spleen, liver, pancreas, and BTCV datasets. Each CT image of the spleen, liver, and pancreas dataset only has a pixel-wise annotation with the corresponding organ position. In contrast, the BTCV dataset has detailed annotations for all organs, including the spleen, liver, and pancreas. Therefore, the BTCV dataset is treated as an out-of-distribution testing dataset, utterly unseen to the other three clients during training and evaluation. The performance on the BTCV dataset provides a good indication of our method's generalization ability. Table 1 presents the details of four datasets.

Brain Tumor MRI datasets. The brain tumor MRI dataset is collected from multi-parametric Magnetic Resonance Imaging scans of brain tumors in the MSD4 dataset. It comprises 484 MRI images from the collaborative institutions in a real-world scenario, where each image consists of four channels: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery

Table 1
Detailed information of Spleen, Liver, Pancreas and BTCV dataset.

Datasets	Image info				Label info		
	Number of images	Slice size [in pixel]	Spacing [*mm*]	Slice thickness [*mm*]	Spleen	Liver	Pancreas
Spleen (Client #1)	41	512	0.71–1.00	1.50–5.00	√	×	×
Liver (Client #2)	131	512	0.56–1.00	0.70–5.00	×	√	×
Pancreas (Client #3)	281	512	0.61–0.98	0.70–7.50	×	×	√
BTCV (out of distribution test)	30	512	0.59–0.98	2.50–5.00	√	√	√

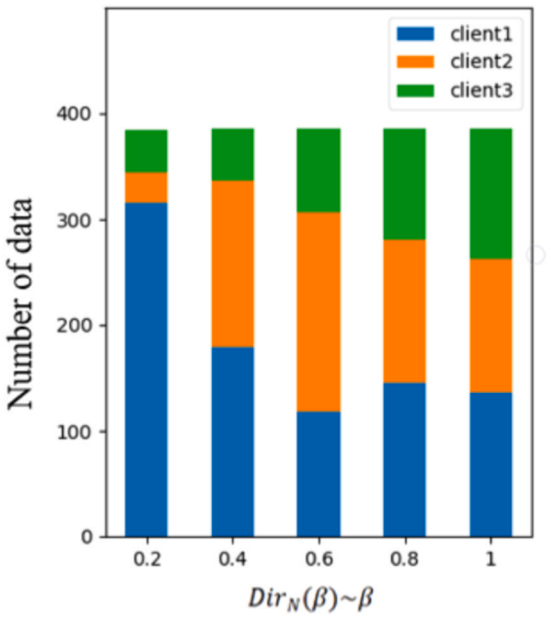

Fig. 3. Training sample allocated to each client at different Dirichlet distribution β.

(FLAIR). The target segmentation categories are split into whole tumor (WT), enhancing tumor (ET), and tumor core (TC).

We utilize a Dirichlet distribution to model statistical heterogeneity by grouping dataset of different sizes. Unlike real-federated data partitions, simulating the heterogeneity distribution of brain tumor datasets can flexibly and efficiently investigate our model performance. We randomly partition the data into three local clients through the Dirichlet distribution $\mathrm{Dir}_{N=3}(\beta)$. A small β indicates a higher degree of data heterogeneity, following similar approaches, we simulate scenarios with IID and different degrees of Non-IID by selecting β values of 0.2, 0.4, 0.6, 0.8 and 1.0.When β is 1 and 0.8, each client has almost the same number of training sample, simulating data homogeneity. When β is less than 0.6, it indicates that each client has been allocated different amounts of data samples. By choosing β values of 0.2. 0.4 and 0.6,we covers IID, normal Non-IID and extreme Non-IID scenarios. The relevant division are illustrated in Fig. 3.

4.2. Implementation details

1) **Self-supervised pre-training setting**: During the pre-processing stage of federated self-supervised pre-training, we apply data augmentation operations to the original images, including random color jittering and random horizontal flipping, followed by normalizing the intensity of each 3D volume using min-max normalization to $[x_1, x_{99}]$, where x_p represents the p th intensity percentile in the 3D volume. The 3D volume is resampled to [178,178,178] for the CT image dataset and [128,128,128] for the Brain Tumor MRI dataset. All methods are implemented using MONAI [34] deep learning framework and deployed in a distribution training system using the Flower framework [35]. ViT3D-B [11] is selected as the backbones for the proposed models. The volumetric images are split into $16 \times 16 \times 16$ patches, which constitute a random mask with a ratio of 75 %. Adams optimization with a weight decay of 0.1 is employed for optimization and the learning rate is set to $1e^{-3}$ with epoch warm-up and decay to$1e^{-5}$with a cosine schedule. Followed by mask self-distillation learning, the weight of the momentum parameter is set to 0.99 and linearly increases to 0.9999. All experiments are conducted using eight NVIDIA 3090 GPU. Considering the influence of communication rounds, a higher number of communication rounds generally enhances model performance. However, the communication cost is higher and its benefits become less significant after a certain number of rounds. In this case, the self-supervised pre-trained model is trained for 25 communication rounds with 30 epochs per round as the default settings. The distribution learning system sets the active client percentage to 100 % for each communication round.
2) **Self-supervised fine-tuning setting**: During federated fine-tuning, we only apply the intensity normalization and shape resampling for input volumetric data. All models are trained based on MONAI and the distribution learning framework Flower. UNETR[1] is chosen as the segmentation model, where the encoder of UNETR is initialized from the pre-trained encoder and trained using a small number of annotation samples. In the fine-tuning, the model is trained for 25 rounds with 30 epochs per round. Adam optimizer is used with a batch size of 8, a learning rate 0.0005, and a cosine schedule.
3) **Evaluation:** During the fine-tuning stage, we evaluate the segmentation performance on Multi-organ CT and Brain Tumor MRI datasets, representing the out-of-distribution test set and data heterogeneity scenario, respectively. The Dice similarity coefficient (DSC) is employed as the evaluation metric.
4) **Baselines:** We present multiple compared baselines to evaluate the proposed self-supervised federated learning, including (1) no pre-training (ViT scratch), where the model trains from random initialization. (2) SimCLR [14], MoCo [15], and BYOL [16] are state-of-the-art self-supervised contrastive learning approaches that incorporate a distillation module for pre-training. We combine these self-supervised methods with FedAvg as federated learning baselines, denoted as FedSimCLR, FedMoCo, and FedBYOL. (3) Fed-MAE [11] is the first federated self-supervised pre-training approach that employs masked image modeling as the self-supervised task. Fed-BEiT [11] is a SOTA federated self-supervised approach in medical image analysis combined with the transformer to facilitate effective presentation. In experiments, we denote the proposed mask self-distillation method as MVE and combine it with personalized federated learning as FedATA.

[1] http://medicaldecathlon.com/.
[2] https://competitions.codalab.org/competitions/17094.
[3] https://www.synapse.org/#!Synapse:syn3193805/wiki/89480.
[4]https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/.

5. Results

Multi-organ CT Segmentation: The quantitative and quality results are presented in Table 2 and Fig. 4. We observe that our method FedATA is more robust than all baselines under both out-of-distribution test set and heterogeneity distribution. First, the advantage of mask self-distillation is pronounced with an improvement of 1.6 % and 0.6 % compared with MAE [8] and BEiT [9] when all self-supervised methods achieve more better performance than random initialization. Next, we observe that the pancreas performs worse than the spleen and liver, with

Table 2
Quantitative result for centralized and distributed learning on BTCV dataset.

Method	Setup	Backbone	Out-of-distribution BTCV DSC %			
			Spleen	Liver	Pancreas	Average
Rand. init.	Centralized	Vit3D-B	81.66	87.62	57.29	75.52
MAE [8]	Centralized	Vit3D-B	82.05	89.20	58.28	76.51
BEiT [9]	Centralized	Vit3D-B	83.71	90.03	59.09	77.61
MVE (ours)	Centralized	Vit3D-B	83.58	90.82	60.35	78.25
FedATA (ours)	Federated	Vit3D-B	**85.65**	**92.54**	**62.18**	**80.12**

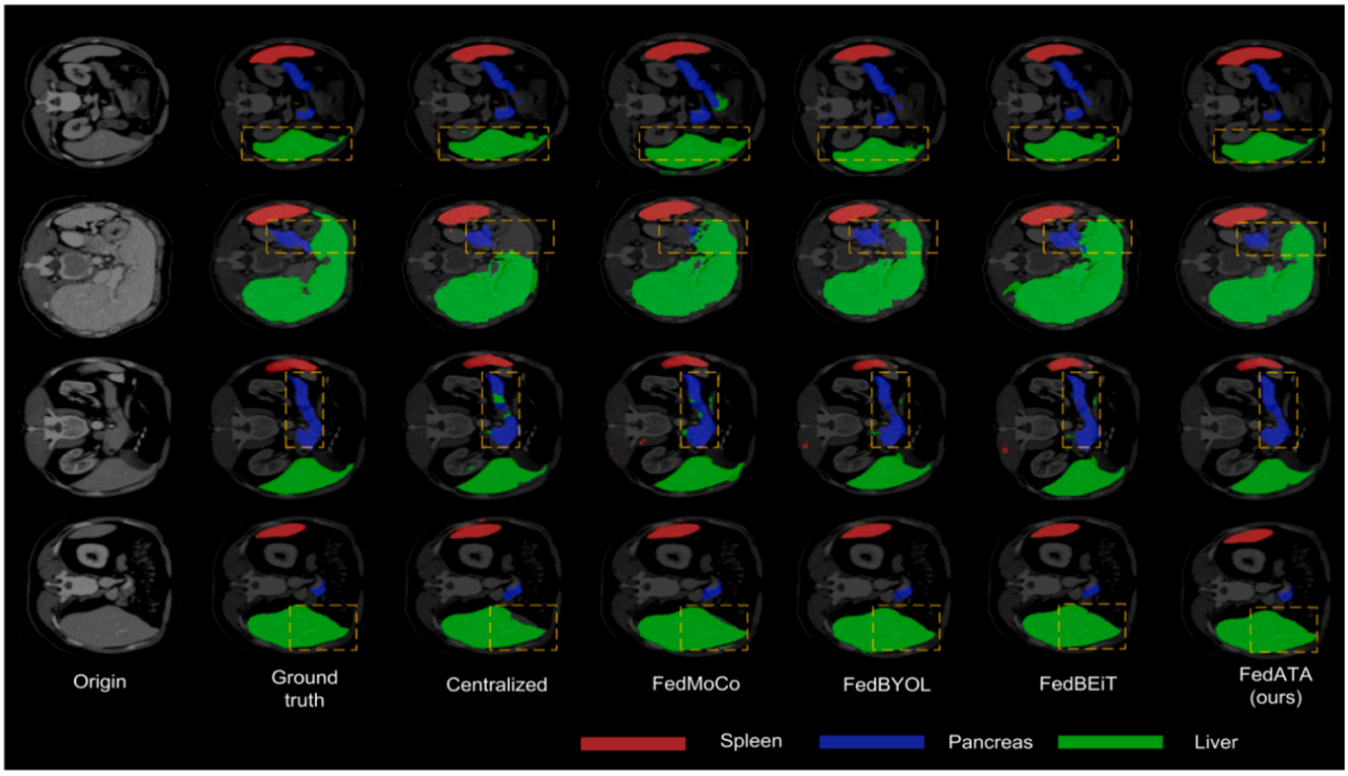

Fig. 4. Visualization of segmentation results on BTCV test dataset. The results are generated from the supervised federated learning fine-tuned model. The proposed method achieves significantly better segmentation performance than other baselines.

Table 3
We compare the proposed method with federated self-supervised pre-training baselines on BTCV dataset.

Method	Distillation	Federated	Out-of-distribution BTCV DSC %			
			Spleen	Liver	Pancreas	Average
FedSimCLR	√	√	82.99	88.21	58.71	76.63
FedMoCo	√	√	83.42	88.52	57.99	76.64
FedBYOL	√	√	83.61	89.37	58.46	77.14
FedMAE	×	√	84.26	89.41	59.39	77.68
FedBEiT	×	√	84.55	90.14	60.11	78.26
FedATA(ours)	√	√	**85.65**	**92.54**	**62.18**	**80.12**

Table 4
Quantitative results for federated learning optimization methods on BTCV dataset and brain tumor dataset.

Federated Learning Method	Out-of-distribution BTCV DSC %				Brain tumor DSC %			
	Spleen	Liver	Pancreas	Average	ET	TC	WT	Average
FedAvg [13]	83.01	90.16	59.88	77.68	73.0	50.5	72.7	65.4
FedBN [31]	83.39	90.62	59.11	77.71	74.7	52.6	72.2	66.5
FedTP [29]	83.70	90.98	60.30	78.33	75.7	52.7	74.0	67.5
FedProx [3]	84.23	91.42	60.69	78.78	76.0	53.0	74.4	67.8
FedPerfix [30]	85.02	92.13	61.06	79.40	76.7	53.6	74.7	68.3
FedATA(ours)	**85.65**	**92.54**	**62.18**	**80.12**	**78.1**	**54.5**	**76.0**	**69.6**

dice scores of 57.29 %, 81.66 %, and 87.62 % respectively. The complex shape and low contrast of the pancreas organ contribute to its lowest dice score, highlighting the challenges and distinctiveness of medical image segmentation. Furthermore, it is worth noting that centralized learning MVE dropped by 1.9 % compared with federated learning FedATA in terms of average dice score, where centralized learning often represents the upper-bound performance of the federated learning models in most cases. Next, we observe that the pancreas performs worse than the spleen and liver, with dice scores of 57.29 %, 81.66 %, and 87.62 % respectively. This observation proves that our method FedATA improves the segmentation performance of the model with limited annotation in a distribution manner.

We further compare our method FedATA with the state-of-the-art self-supervised federated learning methods, specially contrastive learning and masked image modeling. Table 3 presents our method is 3.5 % and 3.0 % higher than FedSimCLR and FedBYOL, which is the 2.5 % and 1.9 % improvement compared to FedMAE and FedBEit. While all of the self-supervised federated learning methods improve the performance compared to centralized learning, our method FedATA can facilitate local fine-tuning with partially labeled data, thus

Table 5
The dice score for localized, centralized, self-supervised federated learning on split-1, split-2, and split-3 (non-IID).

Method		Brain Tumor- static heterogeneity DSC %											
		Split1 (β= 0.2)				Split2 (β= 0.4)				Split3 (β= 0.6)			
		WT	ET	TC	Avg.	ET	TC	WT	Avg.	ET	TC	WT	Avg.
Swin-UNETR [37]	Localized	77.6	45.4	72.4	65.0	76.4	43.1	71.2	63.6	71.6	46.7	68.1	62.2
UNETR++ [36]	learning	77.7	44.0	71.9	64.1	75.9	40.1	71.5	62.4	70.4	47.5	67.0	61.7
UNETR [1]		76.2	6	71.4	63.7	75.1	39.1	71.2	61.8	70.2	45.3	67.5	61.0
TransBTS [38]	Centralized	77.9	57.4	73.5	69.6	78.4	52.2	76.6	69.1	76.1	55.1	73.9	68.3
UNETR++ [36]	learning	78.8	**58.8**	**76.9**	**71.5**	78.1	55.7	75.9	**69.8**	76.3	**55.2**	**74.6**	**68.9**
UNETR [1]		**78.9**	58.5	76.1	71.1	77.8	**56.2**	74.3	69.4	**76.7**	54.7	74.2	68.5
FedSimCLR	Federated	77.6	52.8	74.5	68.4	77.0	48.1	76.2	67.0	76.0	52.4	72.1	66.8
FedMoCo	learning	77.7	53.7	75.8	69.0	77.5	49.6	75.7	67.6	74.2	53.2	74.1	67.2
FedBYO		78.1	52.8	76.5	69.2	77.6	51.0	75.5	68.0	75.6	53.3	73.1	67.3
FedMAE		77.7	54.5	76.1	69.4	77.1	51.7	75.7	68.2	76.2	53.4	73.6	67.9
FedBEiT		77.9	54.8	76.5	69.7	77.6	51.5	76.1	68.4	76.3	53.8	73.5	67.8
FedATA(ours)		78.6	55.9	76.6	70.4	**78.9**	52.5	**76.8**	69.4	76.4	55.1	74.4	68.6

Table 6
The computation and communication cost for complexity analysis.

Method	Computation cost			Communication cost
	Model Params (M)	GFLOPs (G)	Time/iter (s)	Per client (GBits)
FedSimCLR	190.17	48.34	236	79.74
FedMoCo	186.55	47.48	231	78.22
FedBYOL	180.09	45.69	223	75.51
FedMAE	182.07	44.87	226	76.34
FedBEiT	178.58	43.49	222	74.88
FedATA (ours)	**152.03**	**31.19**	**189**	**63.75**

outperforming these baselines. Furthermore, we compared our method to various federated learning optimization methods, e.g., FedAvg [13], FedProx [3], FedTP [29], and FedBN [31]. As shown in Table 4, We observe that these methods can improve performance by 1.72 %, 1.10 %, and 0.65 % compared to FedAvg. However, the gain of 2.44 % from using our method FedATA is significantly larger than the best baseline FedPerfix [30] as novel personalized federated learning.

Brain Tumor segmentation: We conduct further experiments to evaluate the model performance on Brain Tumor datasets. Note that we adapt model training for the brain tumor datasets on localized, centralized, and self-supervised federated learning. Table 5 shows that our method is consistently robust under different levels of data heterogeneity, demonstrating the effectiveness of utilizing partially labeled datasets through FSL. Our method FedATA significantly outperforms the best baseline UNETR++ [36] on localized learning, showing gains of 5.7 %, 7.0 %, and 6.9 %. However, compared to self-supervised centralized learning, our method experiences a slight decline due to the static shift in the training datasets. Compared to UNETR++, our method experiences a drop of 1.1 %, 0.4 %, and 0.3 % on centralized learning, while UNETR++ on localized learning experiences a drop of over 6 %. If a large centralized medical dataset is available, self-supervised learning in a centralized setting may be an excellent alternative to our proposed method.

We further compare our FedATA with the state-of-the-art self-supervised federated learning methods. As shown in the Table 5, FedATA outperforms these baselines by improving 0.7 %, 0.9 %, and 1.2 % compared to FedBEiT, FedMAE, and FedBYOL, respectively. This proves our argument that FedATA is able to effectively extract the desired information from the global model for local fine-tuning, compared to other state-of-the-art self-supervised federated learning baselines, thus achieving superior performances on medical image segmentation. We further observe that FedPerfix improves by 2.9 % compared to the FedAvg baseline through adding the personalized parameters of plugin. Nevertheless, the gain of 4.2 % achieved by our method is more noticeable than using FedPerfix.

Complexity analysis: we present an analysis on the complexity of our FedATA approach in terms of computation and communication cost. The computation cost consists of three components: floating-point operation per second (FLOPS), model parameters (M) and computation time per iter (Second). Table 6 compared the difference between FedATA and other federated self-supervised learning methods. The FedATA has the smallest number of parameters and requires the fewest FLOPS, resulting in faster data transmission between local clients and the server, as well as lower computational demands on local clients. FedATA achieves significant accuracy improvements in a shorter amount of time. In contrast, Fed-SimCL, FedMoCo and FedBYOL

Table 8
Ablation study on the distillation loss format.

Loss	BTCV			Brain tumor		
	Spleen	Liver	Pancreas	WT	ET	TC
baseline	**85.6**	**92.54**	**62.18**	**78.1**	**54.5**	**76.0**
- regulation	83.1	90.92	60.51	76.7	52.2	73.9
+ pixel reg	84.7	91.30	61.33	77.1	51.7	75.7

Table 7
Ablation study on the impact of limited annotation.

Model		Spleen				Liver				Pancreas			
Annotated Ratio (%) of label		20 %	40 %	60 %	80 %	20 %	40 %	60 %	80 %	20 %	40 %	60 %	80 %
Localized learning	Spleen	29.00	42.59	61.20	71.67	-	-	-	-	-	-	-	-
	Liver	-	-	-	-	45.17	59.31	72.68	81.48	-	-	-	-
	Pancreas	-	-	-	-	-	-	-	-	25.51	30.58	42.21	50.81
Distributed learning	FedSimCLR	40.53	61.60	71.54	78.80	54.58	65.94	77.69	83.93	33.28	39.15	50.29	54.84
	FedMoCo	43.50	62.04	73.25	78.25	57.26	72.35	80.26	85.89	35.97	40.84	52.80	55.37
	FedBYOL	42.48	64.98	73.08	78.93	58.47	75.48	81.96	86.68	37.84	42.21	53.15	56.34
	FedMAE	43.43	66.33	73.85	81.42	60.51	76,94	81.48	87.99	37.14	44.64	53.90	58.07
	FedATA(ours)	**49.56**	**68.08**	**77.52**	**83.58**	**65.94**	**79.01**	**86.72**	**89.06**	**39.51**	**48.19**	**56.45**	**60.51**

typically require more time due to the slower process of representation learning through contrastive learning. The table also highlights the communication overhead per client per iteration. FedATA requires lower communication overhead primarily because it only downloads student models in each iteration, whereas other methods have higher overhead due to the need to upload and download the large model.

6. Ablation study

We compare our approach with other alternatives to demonstrate the effectiveness of our model designs, mainly including limited annotation, distillation loss format, and mask ratio on model performance.

1) **Impact of limited annotation:** We examine the impact of limited annotation in the Multi-organ CT dataset and represent our results in Table 7. Firstly, with 20 %, 40 %, 60 %, and 80 % labeled samples in each client, our method FedATA significantly surpasses the localized learning baselines by 4.3 %, 3.8 %, 3.6 %, and 2.9 % respectively. Secondly, the proposed approaches effectively reduce the need for annotations during fine-tuning. For instance, with only 20 % or 40 % annotations per client, our method FedATA achieves significant performance improvement over the top-performing baseline with multiple annotations per client (0.495 vs. 0.425, 0.659 vs. 0.593, and 0.395 vs. 0.30), demonstrating the efficacy of our method in fine-tuning with fewer labeled image.

2) **Distillation loss format:** Unlike previous methods that utilized the cross-entropy (CE) loss to calculate per-pixel distances, we add the regularization term to the L1 loss as distillation loss. This loss aims to minimize the representation distance between the predicted student model and the target of the teacher model. Table 8 demonstrates that the model's performance in both the multi-organ and brain tumor datasets suffered a significant drop when no regularization terms were applied. The reason may be that implementing the regularization term helped to maintain consistency in the feature space between the student network and teacher network, alleviating knowledge forgetting in the student network during training. In this study, we also compared target prediction, which extends the training sample's raw pixel regression. However, adding the pixel regression loss did not lead to an improvement due to a conflict with the regression of deep latent representations produced by the teacher network.

3) **Decoder depth & mask ratio:** Next, we investigate the impact of decoder depth on the student network decoder. As shown in Fig. 5, we find that a single-layer decoder works well for the student network. Increasing the decoder depth appears to worsen performance in brain tumor segmentation. We argue that a too-deep decoder results in an ineffective encoder, relying on a strong decoder to predict the deep latent mask feature. Furthermore, a high mask ratio (75 %) works well in MAE, but a suitable mask ratio for the student network still needs to be explored. In general, predicting latent representation features is more challenging than predicting pixels. While the observations are consistent with MAE results, where an optimal mask ratio tends to yield good results. The reason is that when the mask ratio is similar to MAE pre-training, the student model can express itself to the greatest extent.

4) **Contribution of key components:** we further conduct the ablation for the crucial components in our method, including the exponentially moving average G_e^c, Distillation Loss $L_{con}^f$and Attention Adaptive Aggregation G_a^s. As show in Table 9, removing any of these components leads to performance decreasing across both segmentation tasks. For comparison, we established the baselines M1 to represent our model without Attention Adaptive Aggregation G_d^c. The result indicates that method M1 performs the worst, highlighting the importance of capturing task-specific representations to improve model performance in federated learning. Additionally, we investigated the impact of incorporating a distillation loss, which facilitates knowledge transfer from a larger model to enhance the student model's performance. The results also show that the generalized representation by model M4 is suboptimal when the distillation loss is not included.

5) **Impact of adaptive attention aggregation in FL: In this study, we conducted an examination of the impact of adaptive attention aggregation (ATA) of partial personalized FL and adaptive local aggregation (ALA) of full-model personalized FL. As illustrated in** Fig. 6, we observed that the benefit from the adaptive attention aggregation was notable when compared to adaptive local aggregation (ALA) during federated learning. This indicates that partial self-attention layers contain specific-related information, leading to superior performance compared to full-model personalization while using minimal parameters.

6) **Effect of Communication rate:** The communication rate refers to the number of communication rounds ($T_{total} = T_s + T_f$), where T_s presents the number of communication rounds for pre-training and T_f represents the number of communication rounds for fine-tuning. As shown in Fig. 7, the dice score was gradually increased on the BTCV and Brain Tumor dataset when T_s was set between 15 and 25 iterations. A pronounced decrease was observed for the communication rates exceeding 30 rounds.

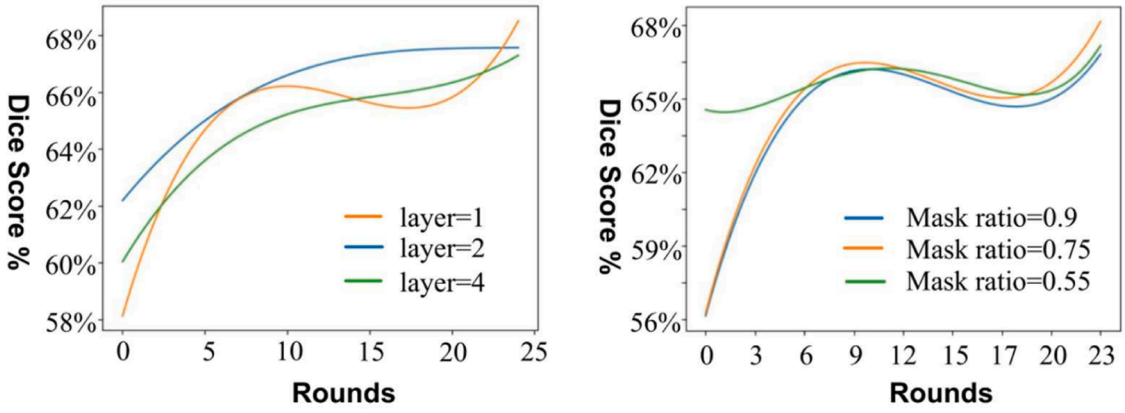

Fig. 5. Ablation study on the decoder depth and mask ratio using Brain Tumor dataset.

Table 9
Ablation study on the impact of limited annotation.

Methods	G_e^c	G_a^s	L_{con}^f	BTCV			Brain tumor		
				Spleen	Liver	Pancreas	ET	TC	WT
M1	√	×	√	83.65	90.83	60.11	73.0	50.9	73.1
M2	×	√	×	84.46	91.15	60.84	74.3	51.5	74.5
M3	√	√	×	84.83	91.44	61.05	75.4	52.4	75.3
M4	×	√	√	85.19	91.69	61.34	76.9	53.8	75.8
FedATA(ours)	√	√	√	**85.60**	**92.54**	**62.18**	**78.1**	**54.5**	**76.0**

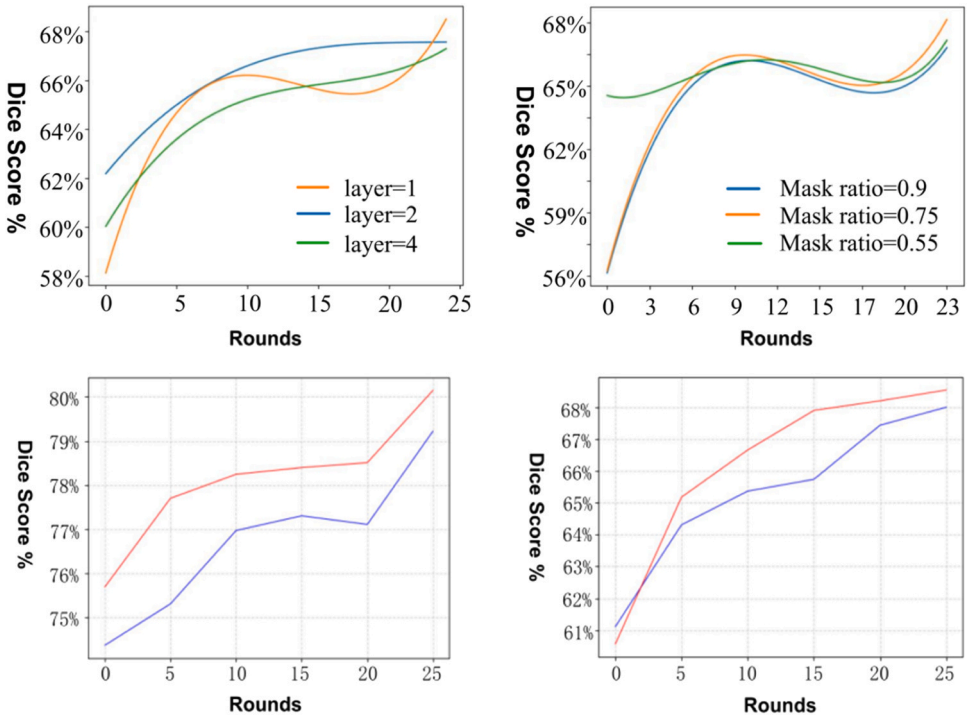

Fig. 6. Ablation study on adaptive attention aggregation using BTCV dataset (left) and Brain Tumor dataset (right).

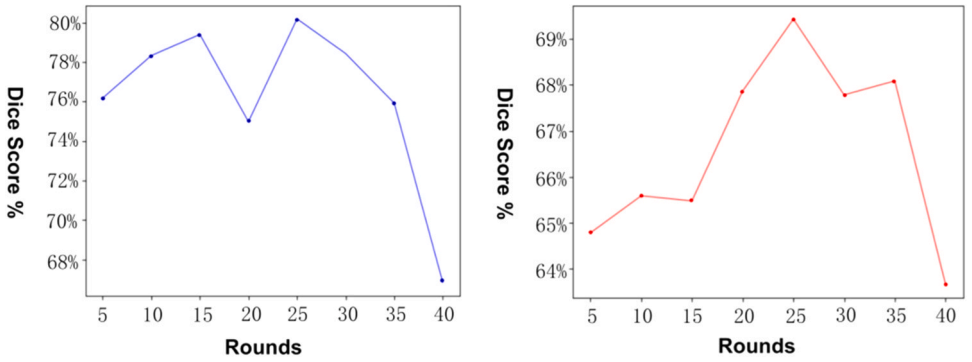

Fig. 7. Ablation study on the number of communication rounds using BTCV dataset (left) and Brain Tumor dataset (right).

7. Conclusion

In this paper, we propose a privacy-persevering and self-supervised federated learning framework for medical image segmentation on decentralized data. This framework employs mask self-distillation to improve the descriptive capability of the learned representation. Furthermore, we utilize adaptive attention aggregation with personalized federated learning to facilitate local fine-tuning under severe data heterogeneity. Our framework exhibits robustness to non-IID data distribution across clients and can effectively learn with limited labeled data. Across the application to the segmentation task, we show that our proposed method outperforms the existing federated self-supervised learning and optimization-based methods under non-IID and label-deficient scenarios.

A potential limitation of this approach is the significant bandwidth usage when applying large foundation models with FedATA for all organ segmentation, particularly when datasets are distributed across multiple clinical sites. The size of these models can range from 110 million to 175 billion parameters, leading to substantial communication and computation costs during parameter updates in the pre-training process. To

mitigate this issue, Low-Rank Adaptation (LoRA) can be integrated into our approach to effectively reduce the number of trainable parameters, which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer.

CRediT authorship contribution statement

Jian Dai: Writing – original draft. **Daoying Geng:** Writing – review & editing. **Huan Liu:** Writing – review & editing, Conceptualization. **Xiao Liu:** Writing – review & editing. **Xing Hu:** Writing – review & editing. **Hao Wu:** Writing – review & editing. **Liheng Yu:** Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daoying Geng reports financial support was provided by the National Nature Science Foundation of China (8237071280). Daoying Geng reports financial support was provided by Shanghai Municipal Commission of Science and Technology (22TS1400900, 23S319041000). Daoying Geng reports financial support was provided by Greater Bay Area Institute of Precision Medicine (Guangzhou) (KCH2310094). Daoying Geng reports financial support was provided by The General Program of Shanghai Natural Science Foundation (22ZR1409500). Hao Wu reports financial support was provided by Young Talents of Shanghai Health Commission (2022YQ043). Hao Wu reports financial support was provided by Medical Engineering Fund of Fudan University (yg2022-2). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Hatamizadeh, Y. Tang, V. Nath, Unetr: Transformers for 3d medical image segmentation, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2022, pp. 574–84.
[2] M.J. Sheller, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (1) (2020) 1–12.
[3] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Res. 2 (2020) 429–450.
[4] G. Kaissis, et al., End-to-end privacy preserving deep learning on multi-institutional medical imaging, Nat. Mach. Intell. 3 (6) (2021) 473–484.
[5] J. Wang, Q. Liu, H. Liang, G. Joshi, H.V. Poor, Tackling the objective inconsistency problem in heterogeneous federated optimization, Adv. Neural Inf. Process Syst. 33 (2020) 7611–7623.
[6] S. Azizi, , Big self-supervised models advance medical image classification, in: Proc IEEE Int Conf Comput Vis, 2021, pp. 3478–88.
[7] X. Li, et al., Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis, IEEE Trans. Med. Imaging 40 (9) (2021) 2284–2294.
[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proc. IEEE Comput. Conf. Comput. Vis. Pattern Recognit, 2022, pp. 16000–9.
[9] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, arXiv 2106 (2021) 08254.
[10] F. Zhang, et al., Federated unsupervised representation learning, Front. Inf. Technol. Electron. Eng. 24 (8) (2023) 1181–1193.
[11] R. Yan, et al., Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging, IEEE Trans. Med. Imaging (2023).
[12] Y. Wu, D. Zeng, Z. Wang, Y. Shi, J. Hu, Distributed contrastive learning for medical image segmentation, Med. Image Anal. 81 (2022) 102564.
[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized dat, Artif. Intell. (2017) 1273–1282.
[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, Proc. Int. Conf. Mach. Learn. (2020) 1597–1607.
[15] J.-B. Grill, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.
[16] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, Adv. Neural Inf. Process. Syst. 33 (2020) 12546–12558.
[17] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, Proc. Mach. Learn. Health (2022) 2–25.
[18] C. You, Y. Zhou, R. Zhao, L. Staib, J.S. Duncan, Simcvd: simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation, IEEE Trans. Med. Imaging 41 (9) (2022) 2228–2237.
[19] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.
[20] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M.M. Balan, K. Brown, Masked image modeling advances 3d medical image analysis, in: IEEE Winter Conf. Appl. Comput. Vis., 2023, pp. 1970–80.
[21] R. Yan, et al., Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging, IEEE Trans. Med. Imaging (2023).
[22] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, P. Prasanna, Self pre-training with masked autoencoders for medical image classification and segmentation, in: Proc IEEE Int Symp Biomed Imaging, 2023, pp. 1–6.
[23] Z. Zhang, Y. Li, B.-S. Shin, Robust medical image colorization with spatial mask-guided generative adversarial network, Bioengineering 9 (12) (2022) 721.
[24] L. Luo, D. Xue, X. Feng, Automatic diabetic retinopathy grading via self-knowledge distillatio, Electronics 9 (9) (2020) 1337.
[25] G. Li, R. Togo, T. Ogawa, M. Haseyama, Self-knowledge distillation based self-supervised learning for covid-19 detection from chest x-ray images, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2022, pp. 1371–5.
[26] S. Park, J. Kim, Y.S. Heo, Semantic segmentation using pixel-wise adaptive label smoothing via self-knowledge distillation for limited labeling data, Sensors 22 (7) (2022) 2623.
[27] J. Liu, B. Li, Z. Luo, Magnetic type classification in sunspot group based on semi-supervised learning and knowledge distillation, in: IEEE Int. Conf. Intell. Comput. Commun. Processing, 2022, pp. 1526–9.
[28] G. Yang, et al., Uncertainty-aware contrastive distillation for incremental semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2567–2581.
[29] H. Li, et al., FedTP: federated learning by transformer personalization, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[30] G. Sun, M. Mendieta, J. Luo, S. Wu, C. Chen, FedPerfix: towards partial model personalization of vision transformers in federated learning, in: Proc IEEE Int Conf Comput Vis, 2023, pp. 4988–98.
[31] X. Li, M. Jiang, X. Zhang, M. Kamp, Q. Dou, Fedbn: Federated learning on non-iid features via local batch normalization, arXiv 2102 (2021) 07623.
[32] C.-M. Feng, Y. Yan, S. Wang, Y. Xu, L. Shao, H. Fu, Specificity-preserving federated learning for MR image reconstruction, IEEE Trans. Med. Imaging (2022).
[33] J. Liu, B. Li, Z. Luo, Magnetic type classification in sunspot group based on semi-supervised learning and knowledge distillation, in: IEEE Int. Conf. Intell. Comput. Commun. Processing, 2022, pp. 1526–9.
[34] M.J. Cardoso, et al., Monai: an open-source framework for deep learning in healthcare, arXiv 2211 (2022) 02701.
[35] D.J. Beutel, T. Topal, A. Mathur, Flower: a friendly federated learning framework, arXiv 14390 (2007) 2020.
[36] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, F.S. Khan, UNETR++: delving into efficient and accurate 3D medical image segmentation, arXiv 2212 (2022) 04497.
[37] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images, Med. Image Comput. Comput. Assist. Interv. (2021) 272–284.
[38] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: multimodal brain tumor segmentation using transformer, Med. Image Comput. Comput. Assist. Interv. (2021) 109–119.

Jian Dai is currently pursuing the Ph.D. degree in electronic information engineering from Fudan university. His current research interests include computer vision, medical image analysis and biomedical engineering.

Hao Wu received the Ph.D. degree from Fudan university. She is currently a Professor at the Huashan Hospital of Fudan University, China. Her current research interests include healthcare information technology, medical image analysis and biomedical engineering.

Huan Liu is currently pursuing the Ph.D. degree in nuclear medicine and medical imaging from the Huashan Hospital of Fudan University. His current research interests include medical image analysis and biomedical engineering.

Liheng Yu received the master's degree in the Academy for Engineering & Technology, Fudan University, Shanghai, China. Her current research interests include medical image analysis and computer vision.

Xing Hu received the Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, in 2016. He is currently a Professor with the School of Optical Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His current research interests include image processing, computer vision, and machine learning.

Xiao Liu received the Ph.D. degree in Beijing Jiaotong University, China, in 2024. His current research interests include medical imaging analysis, computer vision, healthcare information technology and biomedical engineering.

Daoying Geng is currently a Professor at the Academy for Engineering & Technology, Fudan University, Shanghai, China. His current research interests include medical imaging analysis, computer vision, healthcare information technology and biomedical engineering.